DOCUMENT RESUME

ED 051 862 LI 002 922

AUTHOR	Standera, Oldrich
TITLE	COMPENDEX/TEXT-PAC: RETROSPECTIVE SEARCH.
INSTITUTION	Calgary Univ. (Alberta).
PUB DATE	May 71
NOTE	65p.; Information Systems and Services Division Report No. 9
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Automatic Indexing, Computers, Indexes (Locaters), *Indexing, Information Centers, *Information Retrieval, *Information Science, *Information Systems, *Search Strategies
IDENTIFIERS	Canada, CIS, COMPENDEX, Computerized Engineering Index, Current Information Selection, *Text Pac System

ABSTRACT

The Text-Pac System is capable of generating indexes and bulletins to provide a current information service without the selectivity feature. Indexes of the accumulated data base may also be used as a basis for manual retrospective searching. The manual search involves searching computer-prepared indexes from a machine readable data base produced largely by human abstracting and indexing. **Retrospective searching in the Text-Pac System provides computer matching of a machine-readable data base prepared by human abstracting and indexing against questions manually prepared and translated into the system language. The entire record is scanned for occurrence of the question words and the "hits" from this matching are obtained in the form of a computer printout. The logic and search strategy are the same as used for the Current Information Selection. (related reports are available as ED 044 120, and ED 044 121.) (Author/AB)**

THE UNIVERSITY OF CALGARY

DATA CENTRE

COMPENDEX/TEXT-PAC
RETROSPECTIVE SEARCH

by

OLDRICH STANDERA

May, 1971

INFORMATION SYSTEMS AND SERVICES DIVISION

Report No. 9

CONTENTS

0. ILLUSTRATIONS

1. INTRODUCTION

All retrieval services rendered by scientific and technical information services may be divided into current information searches and retrospective searches. The main distinction between these two categories is that only recent information is subject to searching in current searches to keep the users up to date, whereas an accumulated data base is searched in a retrospective service. This difference is, of course, reflected in a somewhat different set-up of programs even though the basic principles of the searching modules are similar.

We have reported our experience gained in implementing the Current Information Selection (Selective Dissemination of Information) in ISSD Report No. 6 (see 2).

In order to inform potential users of the capabilities of the retrospective search module, we prepared the "COMPENDEX* Retro-Search Instructions." These instructions enable any user to submit his request, and a search editor to formulate the request in a language comprehensible to the system (see 4).

The TEXT-PAC** System is capable of generating indexes, too. The reason why we mention them in this conjunction is that they fit into our CIS and retrospective search structure: periodically created indexes and bulletins are a sort of current information service without the selectivity feature. Indexes prepared of the accumulated data base, on the other hand, may be used as a basis of manual retrospective searching. This method of searching will always be indicated where the circumstances warrant it. We could define it as manual searching of computer-prepared

---

*COMPENDEX tapes are the product of and are supplied by the Engineering Index, Incorporated

**TEXT-PAC is an IBM system whose main author is Dr. S. Kaufman, IBM (see 1).

indexes from a machine-readable data base which was produced mostly as a result of manual (human, intellectual) abstracting and indexing (see 6).

Retrospective Searching in the TEXT-PAC System, on the contrary, could be defined as computer matching of a machine-readable data base prepared as a result of manual (human, intellectual) abstracting and indexing, against one or more questions manually prepared and translated into the system language. The "hits" resulting from this matching are obtained in the form of a computer printout. Unlike some other systems, not only the title or key words (subject headings, descriptors, concepts) are searched. The entire record is scanned for the occurrence of the question words and their groupings as indicated by the logical connectors. As the logic and search strategy are essentially the same as used for the Current Information Selection, anyone wishing to obtain more details should refer to our manuals dealing with this topic (see 3, 5).

I wish to express my sincere thanks to Mr. F. T. Dolan for reading and discussing the manuscript and to Mr. S. Nevlud for looking after the smooth running of the tapes as well as some program changes.

2. RETROSPECTIVE-SEARCH SERVICES GENERALLY

In Figure 1 an attempt is made to divide the retrospective search methods into four groups:

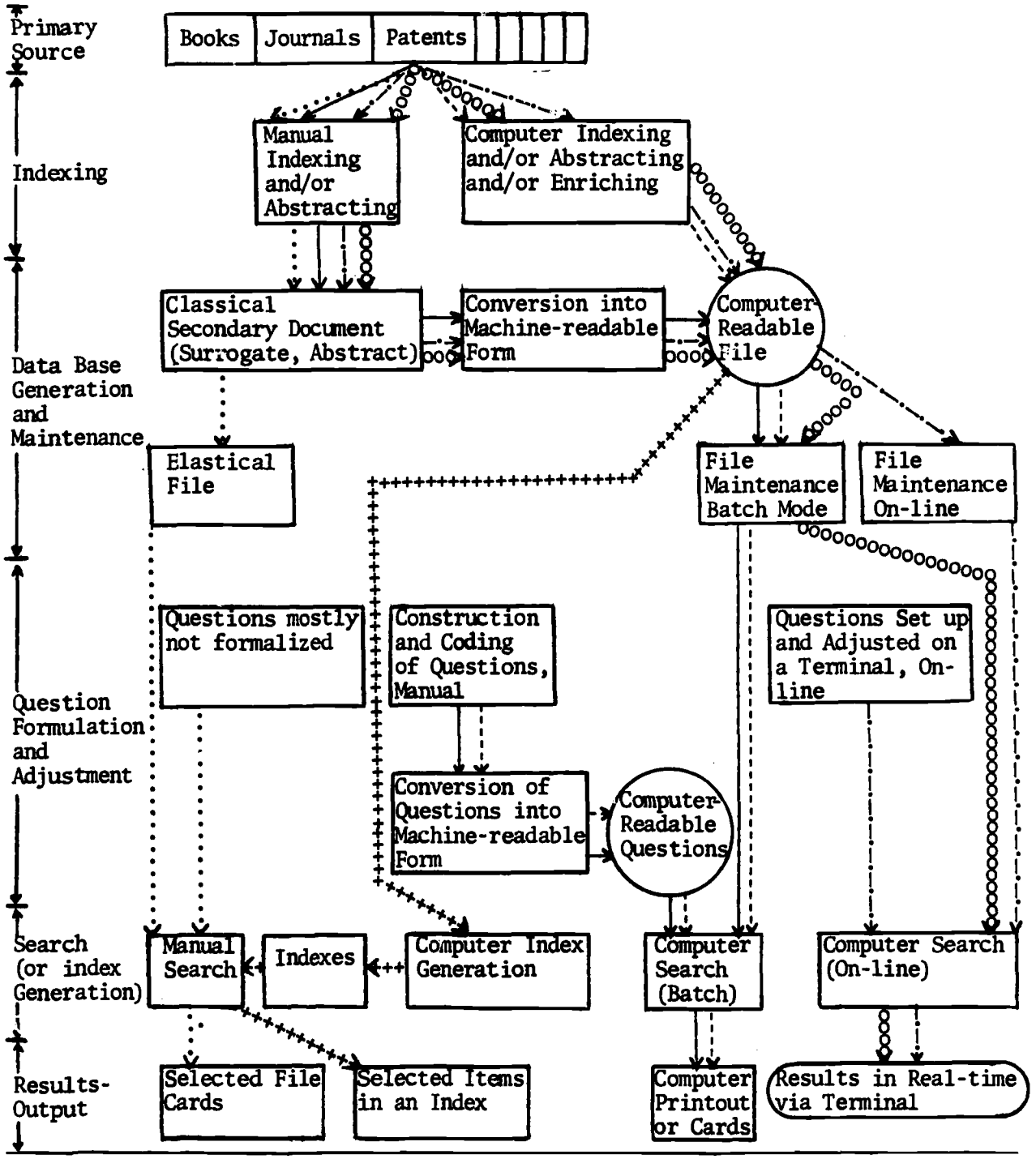
1. Classical Approach with manual indexing (and/or abstracting) and manual search (1 in Figure 1).

2. Manual search in computer-produced indexes (the records prepared either manually or computer-produced items (2 in Figure 1).

3. Computerized methods based on the batch mode, with manual and automatic indexing (3 and 4 in Figure 1).

4. On-line methods (real-time, time-sharing, interactive, conversational) with file maintenance (updating, correcting) on-line or in batch-mode (5 and 6 in Figure 1).

It was not feasible to include all possible modifications and combinations into this simple scheme (e.g., using a terminal question input into batch processing a data base). It has been our intention



Legend:

1) ..... Classical Approach
2) +++++++ Search in Computer-prepared Indexes
3) _____ Batch Mode (Manual Indexing)
4) -------- Batch Mode (Computer Indexing)
5) _._._._._ On-line (File Maintenance On-line)
6) oooooooo On-line (File Maintenance in Batch Mode)

Fig. 1 Retrospective Searching

only to show the position of our COMPENDEX/TEXT-PAC Service in this structure as is illustrated by the full line (3).

It cannot be implied from the above classification that the on-line systems represent something which is unconditionally superior to other methods without considering all other factors involved. Neither can it be deduced that manual methods are inferior under all circumstances. Each method seems to be warranted in a given environment characterized by the level of user requirements, financial considerations, hardware, software, personnel availability, etc.

3. SELECTED DATA ABOUT SOME RETRO-SEARCH SERVICES

The following table presents some data about other retrospective search services and/or systems for such services (Figure 2). As it is very difficult to find data even of a limited degree of comparability, this list is intended to be more of an illustrative sample of what is being done in the field, under certain conditions, than a comparison allowing us to make any general conclusions. Also, the list of the services is by no means complete.

Nevertheless, this table does show the wide gamut of organizations offering retrospective searches including educational, governmental, international, industrial, and research organizations, as well as those institutions specialized in information services. It is evident that data bases of the order of over 1,000,000 records are still considered to be practically manageable, stored on tapes and discs. Various systems handle up to 40 reels in a routine search. Some organizations have limited the number of years back an ordinary search will be performed, or "historical searches" are progressively charged.

Increase of the number of records in the data base over a year is given, with some large services adding as much as 100,000-250,000 records. Here,again, the storage requirements depend strongly on the record size and the useful life of the information contained.

Manual retrospective search from cumulated indexes is still very

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
Patent Search Program. Dow Chemical Company					Batch		$15 per search to qualified personnel	Burroughs B-5500		
ENDS--European Nuclear Documentation System. European Community	850,000 records by the end of 1968	110,000	Boolean		Batch	20 questions against 850,000 records are processed in 45 minutes		360/40		
Science Citation Index. Institute for Scientific Information							$25 minimum		IBM compatible	
Experimental Collection and Retrieval in the Information Sciences. Lehigh University	3,000 documents	900			Batch: (author file, thesaurus searching) On-line: reference retrieval		Remote query: $50/hr. for CPU time; $27/hr. for Datanet time	GE-225 Datanet for remote query		
Project TIP. MIT--Massachusetts Institute of Technology	60,000 records (30,000 of them on disk)				On-line			IBM 7094+ 150 remote consoles		

Continued

(continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
Mechanical Properties Test Data of Structural Materials. Belfour Stulen, Inc.							Industrial users <$50 max. $75 per search	360/25	FORTRAN COBOL ALC RPG	
University of Notre Dame, Radiation Chemistry Data Centre			AND, OR NOT		Batch		No charge at present. In future, computer time, postage, reproduction may be charged.	Univac 418, 1004	Assembly language ART	
On-line Mathematical Citation Index. University of California, Los Angeles	30,000 citations on punched cards and tape	60,000			Batch, eventually On-line			360/85	Assembly language	
Scientific Russian Text. The Rand Corporation	20 reels 7 track 556 or 800 BPI 30,000,000 words							IBM 7044		

Continued

(continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
Computer Tape for Searching the Gas Chromatographie. Preston Technical Abstracts Company			AND, OR				Price References $ 60 1-10 100 11-99 175 100-249 250 250-499 375 500-749 500 750-999	360/65	Assembler	
Union List of Scientific Serials in Canadian Libraries. National Science Library	Data on 2 tapes (2400 9H)						Selective runs, $300/hr. of computing time	360/50	Assembler	
NASA/STAR IAA C-STAR	600,000 citations in cumulative indexes or on tapes	100,000			Manual from cumulative indexes					
NASA/	600,000 citations on 35 tapes	100,000		1 week	Batch	6-8 hrs. whole file	$60-$80 per search (preliminary figures) cost		FORTRAN Assembler	12,000 (30 centres)

Continued

(continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
NASA/"RECON"	600,000 citations on 35 tapes	100,000			On-line 21 terminals; 50-70 terminals are planned	30-50 mins.	$10/search (preliminary study) cost	360/50 512K	125K	
North American Rockwell Corporation				1 week (40-50 questions) 200 questions can be handled economically	Batch		Restricted to company employees only			2,000-2,500
DATRIX/Xerox Corporation, University Microfilm, Inc. (bibliography of dissertations)	126,000 (bibliographical references)				Batch		$5 for the first 10 references and 10¢ for each additional	360/40 with 2314 disc facility		

Continued

(continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
KASC/University of Pittsburgh	7 magnetic tapes		Boolean		Batch		$275 for selected listing with abstracts retrospective from 1962 plus 12 monthly selected listings	IBM 7090		
Sandia Corporation. Book and Report Cataloging	12 reels (800 BPI)							CDC 3600 64K	COBOL 40 programs	
INTREDIS/U.S. Forest Service							$20/search		FORTRAN	
Catalog on the Library of the Pacific Southwest. Forest and Range Experiment Station U.S. Forest Service			Boolean				$30 max./ search	360 and CDC 6400	FORTRAN IV and MAP	
Bibliographical References in the Atmospheric Sciences. National Center for Atmospheric Research	20 reels				Batch			CDC 3800 CDC 6600	FORTRAN	

Continued

continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
Historical Weather Data. U.S. National Weather Records Center	22,000 reels (2,400 feet)							Honeywell 1200 RCA Spectra 70/45		
NARDIS/U.S. Naval Ship Research and Development Center	18 reels							UNIVAC LARC		
BIOSIS--Bio-Sciences Information Service Philadelphia	1,333,000 documents 21 mill. key words (postings)	250,000		6 working days	Batch	72 sec/question	$150/search			
Chemical Titles/ University of Alberta, Edmonton	7 years of Chem. Titles= 12 tapes 2400'; 800 BPI 1,000,000 titles Dictionary= 340,000 different words=300K bytes		AND, OR NOT		Batch	80,000 titles 2.6 minutes (1.5 min. is system overhead) 9 sec./question	$15 registration $2 per item 5¢ per hit	360/67	100K	

Continued

(continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
SIE--Science Information Exchange. Smithsonian Institute	>500,000	100,000 (useful life less than 1 year)					$40/question $30 each additional Historical Search 1949--2 years prior to the current: $75/question $60 each additional	360/30	COBOL	55,000
University of Pittsburgh			AND, OR NOT		On-line	1,100 documents in 5.3 sec. (1 document=25 card images)		360/50	16K	
MEDLARS/National Library of Medicine	More than 1,000,000 citations on 39 reels searches 3 years back	200 000		3 weeks	Batch (on-line under development)		Free to health professional personnel	Honeywell 800-200	Assembly language	10,000 in U.S.A.

Continued

(Continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
ORBIT II delivered by SDC--System Development Corporation	Up to 1,000,000 records 48 million words; max. size of a unit record 7,100 char.				On-line access for 10-150 users at a time. Batch input. Batch or on-line update.	1/10 sec. a typical search		360/40	256K	
Uniterm Index for U.S. Chemical and Chemically-Related Patents. IFI/Plenum Data Corporation, Washington	Past 20 years	420,000 (six bi-monthly at 70,000)	Boolean				$150/search $65 for 50 searches	IBM 360/OS 256K	FORTRAN ALC	
Metals Abstracts Index Data Base. Amer. Soc. for Metals	From 1966	23,000					$250 for complete file search			

Continued

(continued)

Service, Organization/Parameters	Size of Files	Rate of Increase rec/year	Logic	Turn-around Time	Manual Batch On-Line	Time Considerations	Cost/Price	Hardware	Software Core	Number of Questions per year
GEO REF--Amer. Geological Institute, Washington	Past 3.5 years	36,000-48,000					$10 per query per 50 items retrieved			

Fig. 2 Survey of Retrospective-Search Services

popular. Although batch-searching is more widespread, on-line methods are gaining ground, some of them in developmental, others in pilot-plant and full production stages. Some data bases operated under a batch mode are being transferred to an on-line conversational (interactive) time-sharing mode. There may be even hybrid systems where some record fields are searched in batch mode whereas others are on-line searched. Some operations, such as updating, may be done both in a batch or on-line mode. Up to 150 users may have simultaneous access to the file.

The turnaround time has a direct bearing on the mode employed. Whereas 30 seconds seems to be excessive in a conversation mode, a turnaround time of a week is quite common with a batch mode and may extend even to several weeks, depending on the urgency of demands and technical circumstances. The number of questions processed in a run and the turnaround time vary considerably among services; one system can handle 200 questions economically. In batch mode it should be noted that there are three kinds of possible operation: local batch; remote batch (terminal, batch processing, terminal); and deferred batch (terminal, batch processing, peripheral equipment).

The number of questions posed to the system varies widely from service to service. One central service claims to be asked 55,000 questions per year, other well-established services operating in several branches and covering a vast subject field, report over 10,000 questions a year, whereas a big firm with a restricted distribution of output answered 2,000 questions a year. It appears that 300-500 questions per year for a small information centre might be quite justified.

Search-time data are most difficult to compare as they depend on the hardware, software, number of questions and their complexity, the search strategy used, and the size of the data base. Accordingly, the search time for one question was reported to be 9, 72, 135 seconds in three different systems searching 80,000 and 1,333,000 and 850,000 records, respectively. Conversational mode search times are recorded in seconds.

Only in rare cases is retrospective search offered free, usually free service is restricted to staff. Some charges are stated as a lump sum, or a minimum or maximum amount, which could incur. In some cases, the fee is calculated depending on the number of references found. Sometimes a basic charge is set for a certain number of hits and additional hits are extra. Basic fee and question terms and hits may be the basis of the price. Additional questions are sometimes allowed at discount prices. Occasionally computer time only is charged. As may be seen, the pricing policies are very different and reflect the actual costs to a very limited degree. In most cases the operation of a service is subsidized in some way or other.

4. TEXT-PAC RETRO-SEARCH MODULE AND COMPENDEX-TAPE SERVICE

4.1

The complete documentation of the TEXT-PAC software may be found in (1). This system allows the full text of documents to be searched. The programs are in Basic Assembler Language (BAL) and are designed for the IBM's OS/360 (MVT or MFT). The required configuration comprises the system 360 and needs 180K core memory, a card reader, a printer, four 9-track tape drives, and one DASD (e.g., scratch disk as temporary storage).

COMPENDEX is supplied on 9-track tapes 800 BPI in EBCDIC. Tape length is 1,200 feet. It is delivered monthly and contains over 5,000 records. Records are variable length, unblocked, maximum length 8,004 bytes. The input format is TEXT-PAC 360 Condensed Text. More information about the tapes may be obtained from (13).

Each record is classified by Main Subject Headings and Subheadings which are listed in (11). Another access point to the records represents the CAL (Card-A-Lert codes) described in (12).

Publications which are abstracted and indexed for COMPENDEX are listed in (10) together with the type of coverage: complete; partial; or monitored.

The data base (115,000 records) is at present contained on 12 magnetic tapes, or one tape accommodates nearly 10,000 records. The yearly growth is expected to be 60,000-70,000 records, or 6 to 7 tapes.

4.2 Some Limitations in the Retrospective Search

1. Maximum of 200 answers to any question unless otherwise specified (9999 possible) in the field 'Maximum hit count." (See also 15.)

2. Match criterion 01-19.

3. Only one memory load of questions can be processed at a time. If there are any left, another run will be necessary.

4. The maximum number of connected logical symbols (A1, A2 . . .) is 15.

5. More than three levels of back referencing is not permitted. (See 3)

6. Question words and logical symbols must not be mixed in a concept or search expression.

7. A logical symbol must not be referred to more than 15 times in one question.

8. A maximum of 9 continuation cards may be used in a concept or search expression.

9. Any question may be defined by a maximum of 99 cards.

10. Maximum word length in a question is 40 characters.

11. You may specify up to 7 print controls in the CONTROL.

12. All of the specified logical symbols (A1, A2 . . .)must be used in the search expressions inside any question.

13. A maximum of 15 words may be connected by "AND."

14. If the statistical option was requested, a list of up to 20 words causing a hit is printed for each document.

15. "Retrospective Text Sort" can process hits up to the maximum of 6,000. A larger number of hits would necessitate using IBM 360/OS Sort program.

5. STATISTICAL OPTION

As we have already mentioned in our COMPENDEX Retro-Search Instructions (4) the user can obtain statistical data indicating which of the logic (words and logic connectors) has been responsible for the hits, if any were accomplished. This option is specified on the Reader card (column 9) at the time a question is coded.

The statistical printout (or trigger cards) could be used, theoretically, to one or both of these objectives:

1. To decide what documents hit by the question should be printed. The trigger cards would make it possible. However, it seems to us that a responsible decision in this respect cannot be made with only trigger cards and/or statistical printout at hand. This would necessitate checking over the pertinent abstract in the Edit print which would have to be printed at an extra cost. Checking the printed answers is less time consuming and, therefore, the better alternative.

2. The statistical data about the hit logic provide the means for improving a profile. In this connection it should be stated that the statistical feature being described seems to be more appropriate in the CIS mode, where the profile is of a semi-permanent nature and thus has to be corrected continually on the basis of user's feedback. We can, of course, modify a retrospective question in the event that there are either too many or too few answers.

(a) In the first case, we can make the most prolific search expressions more selective, we can omit ambiguous expressions false drops. We can leave out expressions having no response in the searched data base, so the next search will be faster.

(b) In the second case, we will loosen the question to be more responsive and leave out or modify expressions giving no hits.

Then we can resubmit the question. In any of the above cases, we need the printout, as only by referencing the abstracts can we use the statistical printout in adjusting questions. The reason is that from the statistical printout alone we cannot conclude whether or not the document is relevant.

Figure 3 and Figure 4 illustrate what the three programs do for the user depending on his option. Figure 5 shows the format of the statistical printout. The format of the trigger cards is much the same.

The statistical printout is a valuable tool designed to modify both profiles and questions, whereas the use of trigger cards without studying the pertinent abstracts seems to offer little help. It is more convenient to study the statistical printout and the answers, and modify the question accordingly.

6. TIME AND COST

6.1 Retro-Search Programs

The programs involved in the Retrospective Search (non-statistical) are: Retro-Memory Load; Retro-Search; Retro-Text Expansion; Retro-Text Sort; Retro-Print.

The Retro-Search Program is by far the most time consuming, the Retro Memory Load and the Retro-Text Expansion are negligible even with 100 questions and 60,000 records. The Retro-Text Sort and the Retro-Text Print are worth consideration only with a higher number of questions and records.

In order to ascertain the effect of the number of questions, we have taken a data-base of 60,000 records which resulted from merging of individual monthly tapes, and determined the CPU times of the programs named above. Also, the number of data sets, memory region, and I/O waits are shown where available, for 1, 2, 3, 5, 10, 20, 30, 40, 50, 60, 70, 80, and 100 questions, with 12 hits per question (see the tables below, Figure 6).

To show the relationship between the CPU times and the number of records, we conducted a search for 10 questions against a data base consisting of 5,000; 10,000; 20,000; 40,000; 60,000; and 80,000 records. The results are illustrated in the tables below (Figure 7).

It has been shown that the CPU time of the search programs is influenced by the number of questions (after the initial sharp increase,

Program	"S" wanted	"S" not wanted
Retro-Statistical	Printout contains the card images of the question and (1) either statistical data (Fig. 5) if any hits were achieved, and trigger cards if not circumvented (2) or "No hits" message	Printout contains question, card-images and the number of hits
Retro-Print First Pass	Printout contains edited question and "No answers for this question" message	Printout contains edited questions and the found documents
Retro-Print Second Pass	Printout contains edited question and answers. (These may be monitored by the trigger cards, either all or part of them, either positively or negatively. At least one header card must be present and may, in addition, change the title or print-controls to be printed.)	

Fig. 3 Statistical Option: What the Programs Do

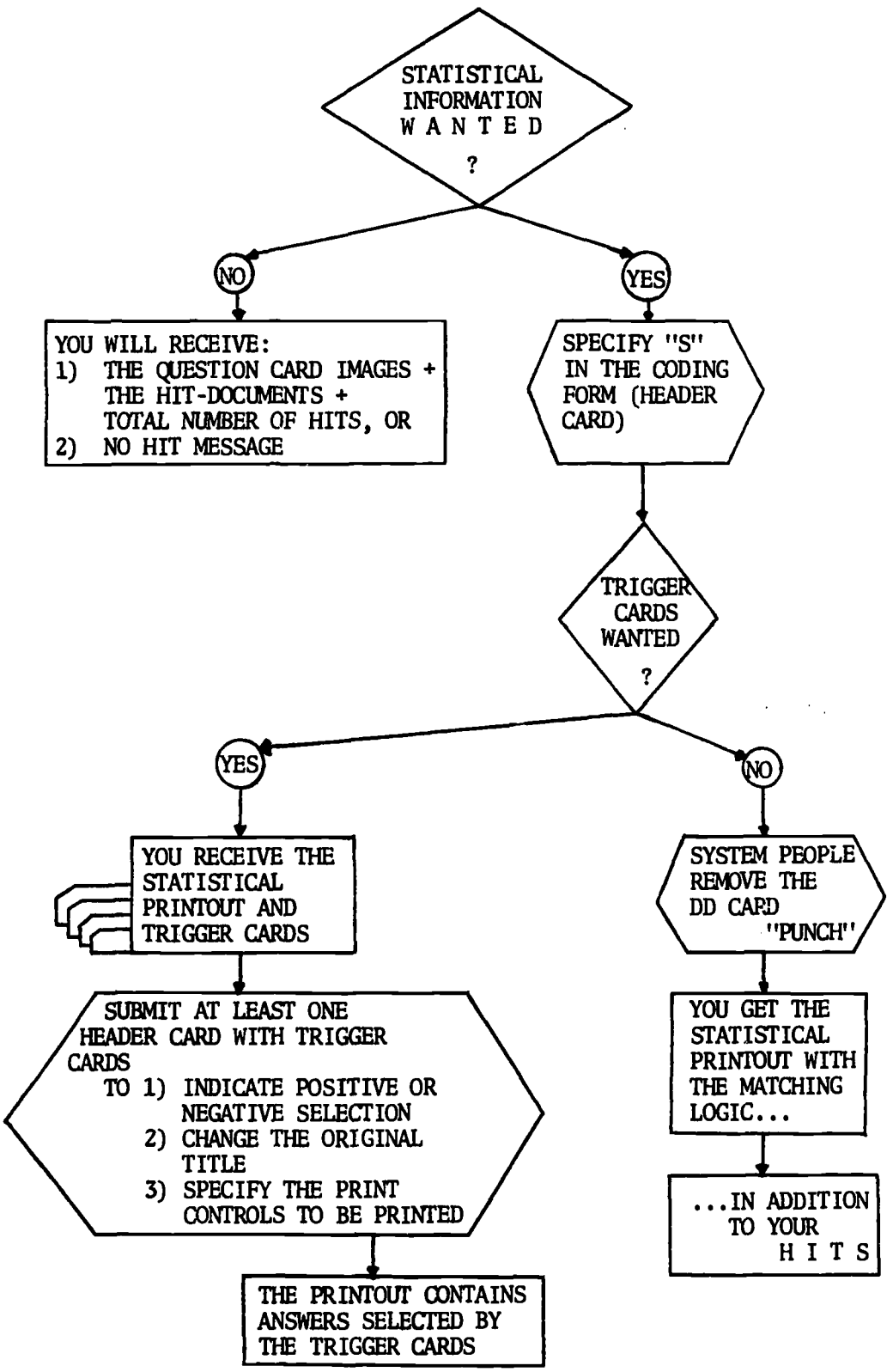


Fig. 4 Statistical Option: Decision making

Question No.	Answer Number	Required	Found	Word Matches Found $(And) *(First Word - Adjacency) =(First Word - With) & (First Word - With/And) + (First Word - Adjacency/And)	
R00005	EIX69X100165	1	1	$Dissemination	
R00005	EIX69X120852	1	1	$Dissemination	
R00005	EIX69X122878	1	1	*Current	Awareness
R00005	EIX69X123060	1	1	$Dissemination	
R00005	EIX69X123157	1	2	*Current	Awareness
				$Dissemination	
R00005	EIX69X123173	1	1	$Dissemination	
R00005	EIX70X013031	1	1	$Dissemination	
R00005	EIX70X020001	1	1	$Dissemination	
R00005	EIX70X033661	1	2	$Process	
				=Retrospective	Search
R00005	EIX70X042929	1	1	=Retrospective	Search
R00005	EIX70X054959	1	1	$Dissemination	
R00005	EIX70X070225	1	1	$Running	Text
R00005	EIX70X115750	1	2	*Current	Awareness
				=Retrospective	Search
R00005	EIX70X123065	1	1	*Current	Awareness
R00005	EIX70X123149	1	1	*Current	Awareness
R00005	EIX70X125775	1	1	$Dissemination	
R00005	EIX70X127755	1	1	$Dissemination	
	17 Hits				

Fig. 5 Statistical Printout: Format

1 Question (60,000 Records, 12 Hits/Question)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time CPU
221	Sorted Question Diagnostic	5			
222	Retro-Memory Load	11		58	4 sec.
223	Retro-Search	7		48	8 min. 23 sec.
227	Retro-Text Expansion	4		50	1 sec.
228	Retro-Text Sort	5		72	2 sec.
229	Retro-Print	5		52	3 sec.
					8 min. 33 sec.

2 Questions (60,000 Records, 12 Hits/Question)

221	Sorted Question Diagnostic	5		52	1 sec.
222	Retro-Memory Load	11		58	3 sec.
223	Retro-Search	7		48	10 min. 14 sec.
227	Retro-Text Expansion	4		50	1 sec.
228	Retro-Text Sort	5		72	2 sec.
229	Retro-Print	5		52	4 sec.
					10 min. 25 sec.

Continued

3 Questions (60,000 Records, 12 Hits/Question)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time CPU
221	Sorted Question Diagnostic	5		52	1 sec.
222	Retro-Memory Load	11		58	3 sec.
223	Retro-Search	7		48	12 min. 23 sec.
227	Retro-Text Expansion	4		50	1 sec.
228	Retro-Text Sort	5		72	2 sec.
229	Retro-Print	5		52	5 sec.
					12 min. 35 sec.

5 Questions (60,000 Records, 12 Hits/Question)

221	Sorted Question Diagnostic	5		52	2 sec.
222	Retro-Memory Load	11	*90	58	4 sec.
223	Retro-Search	7	*31,000	50	16 min. 48 sec.
227	Retro-Text Expansion	4	*15	50	1 sec.
228	Retro-Text Sort	5	*300	72	3 sec.
229	Retro-Print	5	*1,700	52	7 sec.
			*Estimate		17 min. 5 sec.

Continued

10 Questions (60,000 Records, 12 Hits/Question)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time CPU
221	Sorted Question Diagnostic	5		52	2 sec.
222	Retro-Memory Load	11		58	4 sec.
223	Retro-Search	7		54	27 min. 26 sec.
227	Retro-Text Expansion	4		50	1 sec.
228	Retro-Text Sort	5		72	6 sec.
229	Retro-Print	5		52	12 sec.
					27 min. 51 sec.

20 Questions (60,000 Records, 12 Hits/Question)

221	Sorted Question Diagnostic				
222	Retro-Memory Load	11	366	58	6 sec.
223	Retro-Search	7	31,242	60	45 min. 37 sec.
227	Retro-Text Expansion	4	66	50	1 sec.
228	Retro-Text Sort	5	1,122	72	9 sec.
229	Retro-Print	5	6,736	52	27 sec.
					46 min. 20 sec.

Continued

30 Questions (60,000 Records, 12 Hits/Question)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time CPU
221	Sorted Question Diagnostic				
222	Retro-Memory Load	11	538	58	8 sec.
223	Retro-Search	7	31,318	68	66 min. 53 sec.
227	Retro-Text Expansion	4	96	50	3 sec.
228	Retro-Text Sort	5	1,682	72	17 sec.
229	Retro-Print	5	10,101	52	37 sec.
					67 min. 57 sec.

40 Questions (60,000 Records, 12 Hits/Question)

221	Sorted Question Diagnostic				
222	Retro-Memory Load	11	715	58	11 sec.
223	Retro-Search	7	31,396	74	79 min. 39 sec.
227	Retro-Text Expansion	4	126	50	2 sec.
228	Retro-Text Sort	5	2,242	72	20 sec.
229	Retro-Print	5	13,280	52	48 sec.
					81 min.

Continued

50 Questions (60,000 Records, 12 Hits/Question)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time CPU
221	Sorted Question Diagnostic	5		52	6 sec.
222	Retro-Memory Load	11	*900	64	13 sec.
223	Retro-Search	7	*31,500	80	107 min. 51 sec.
227	Retro-Text Expansion	4	*150	50	3 sec.
228	Retro-Text Sort	5	*2,800	72	29 sec.
229	Retro-Print	5	*16,000	52	4 sec.
			*Estimate		108 min. 46 sec.

60 Questions (60,000 Records, 12 Hits/Question)

221	Sorted Question Diagnostic				
222	Retro-Memory Load	11	1,067	68	15 sec.
223	Retro-Search	7	31,550	86	121 min. 57 sec.
227	Retro-Text Expansion	4	185	50	3 sec.
228	Retro-Text Sort	5	3,360	72	35 sec.
229	Retro-Print	5	20,190	52	1 min. 12 sec.
					124 min. 2 sec.

Continued

70 Questions (60,000 Records, 12 Hits/Question)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time CPU
221	Sorted Question Diagnostic				
222	Retro-Memory Load	11	1,244	74	18 sec.
223	Retro-Search	7	31,628	92	143 min. 19 sec.
227	Retro-Text Expansion	4	214	50	4 sec.
228	Retro-Text Sort	5	3,919	72	53 sec.
229	Retro-Print	5	23,561	52	1 min. 31 sec.
					146 min. 5 sec.

80 Questions (60,000 Records, 12 Hits/Question)

221	Sorted Question Diagnostic				
222	Retro-Memory Load	11	1,419	80	22 sec.
223	Retro-Search	7	31,705	100	154 min. 42 sec.
227	Retro-Text Expansion	4	244	50	5 sec.
228	Retro-Text Sort	5	4,478	72	57 sec.
229	Retro-Print	5	26,926	52	1 min. 39 sec.
					157 min. 45 sec.

Continued

100 Questions (60,000 Records, 12 Hits/Question)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time CPU
221	Sorted Question Diagnostic	5		52	9 sec.
222	Retro-Memory Load	11		94	32 sec.
223	Retro-Search	7		114	222 min. 38 sec.
227	Retro-Text Expansion	4		50	5 sec.
228	Retro-Text Sort	5		72	1 min. 29 sec.
229	Retro-Print	5		52	2 min. 10 sec.
					227 min. 3 sec.

Fig. 6 Varying Number of Questions

5,000 Records (10 Questions)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time (CPU)
221	Sorted Question Diagnostic	5	193	52	2 sec.
222	Retro-Memory Load	11	228	58	6 sec.
223	Retro-Search	7	2,483	56	3 min. 2 sec.
227	Retro-Text Expansion	4	14	50	1 sec.
228	Retro-Text Sort	5	223	72	3 sec.
229	Retro-Print	5	1,363	52	7 sec.
					3 min. 21 sec.

10,000 Records (10 Questions)

221	Sorted Question Diagnostic	5	193	52	2 sec.
222	Retro-Memory Load	11	228	58	5 sec.
223	Retro-Search	7	4,911	56	5 min. 25 sec.
227	Retro-Text Expansion	4	20	50	1 sec.
228	Retro-Text Sort	5	313	72	3 sec.
229	Retro-Print	5	1,918	52	8 sec.
					5 min. 44 sec.

Continued

20,000 Records (10 Questions)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time (CPU)
221	Sorted Question Diagnostic	5	193	52	2 sec.
222	Retro-Memory Load	11	228	58	4 sec.
223	Retro-Search	7	10,019	56	11 min. 32 sec.
227	Retro-Text Expansion	4	51	50	2 sec.
228	Retro-Text Sort	5	969	72	9 sec.
229	Retro-Print	5	4,966	52	11 sec.
					12 min.

40,000 Records (10 Questions)

221	Sorted Question Diagnostic	5	193	52	2 sec.
222	Retro-Memory Load	11	228	58	4 sec.
223	Retro-Search	7	20,342	56	22 min. 15 sec.
227	Retro-Text Expansion	4	126	50	2 sec.
228	Retro-Text Sort	5	2,384	72	21 sec.
229	Retro-Print	5	12,133	52	43 sec.
					23 min. 27 sec.

Continued

60,000 Records (10 Questions)

TRC No.	Program	No. of Files	I/O Waits	Region (K)	Step Time (CPU)
221	Sorted Question Diagnostic	5	193	52	2 sec.
222	Retro-Memory Load	11	228	56	4 sec.
223	Retro-Search	7	30,685	50	33 min. 22 sec.
227	Retro-Text Expansion	4	202	50	4 sec.
228	Retro-Text Sort	5	3,783	72	47 sec.
229	Retro-Print	5	19,504		1 min. 18 sec.
					35 min. 37 sec.

80,000 Records (10 Questions)

221	Sorted Question Diagnostic	5	193	52	2 sec.
222	Retro-Memory Load	11	228	58	5 sec.
223	Retro-Search	7	41,241	56	45 min. 1 sec.
227	Retro-Text Expansion	4	245	50	4 sec.
228	Retro-Text Sort	5	4,601	72	57 sec.
229	Retro-Print	5	23,683	52	1 min. 45 sec.
					47 min. 54 sec.

Fig. 7 Varying Data Base

directly proportional), by the number of data-base records (directly proportional), and by the number of hits. We have not examined the impact of the number of hits as they can be monitored only indirectly and they vary from question to question. The relationship "CPU time to number of questions" is illustrated in Figure 8. The relationship "CPU time to number of records" is depicted in Figure 9 and Figure 10. In the former case, the number of hits per question was kept constant (12 hits per question); in the latter case, of course, the number of hits per question was increasing with the size of the data base. In the "CPU time per number of records" chart, the effect of looser questions on the search time is clear: the CPU time for 10 questions and 60,000 records equals 35.5 minutes, whereas in the chart "CPU time per number of questions" the CPU time for 10 questions and 60,000 records is less than 28 minutes. This difference reflects the different number of hits (for each of the questions and for all of them, as they are identical) brought about by the looser question structure in the former case (and, therefore, a higher number of hits) and the more selective structure in the latter case.

While we cannot monitor the size of the searched data base as this is determined by users themselves at the time they submit the question, we can to a certain degree control the size of a batch of questions processed each time. An urgent question, of course, would be run anytime, regardless of cost. For this reason we have examined the CPU time per one question when running batches of various size. We have found that one question requires as much as 8.5 minutes of the CPU time to complete the search programs, whereas with a 40-question batch only 2 minutes per question are needed (these figures were obtained in searching 60,000 records with 12 hits per question). No considerable rise of this time was found up to 100 questions (see Figure 11). For data bases containing a higher number of records, the CPU time required per question will be higher in batches of any size, but the form of the curve will remain unchanged.

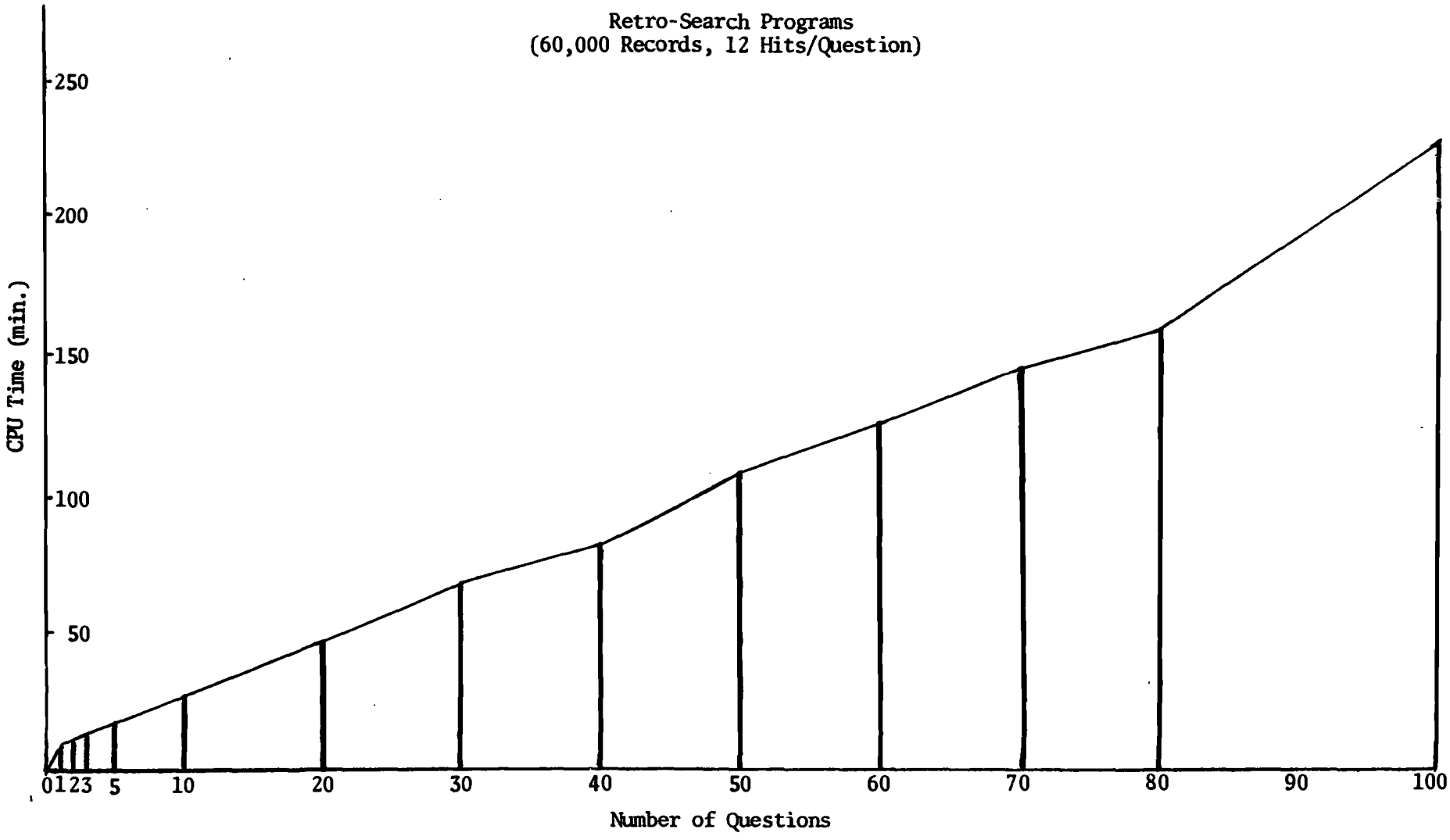


Fig. 8 Number of Questions vs. CPU Time

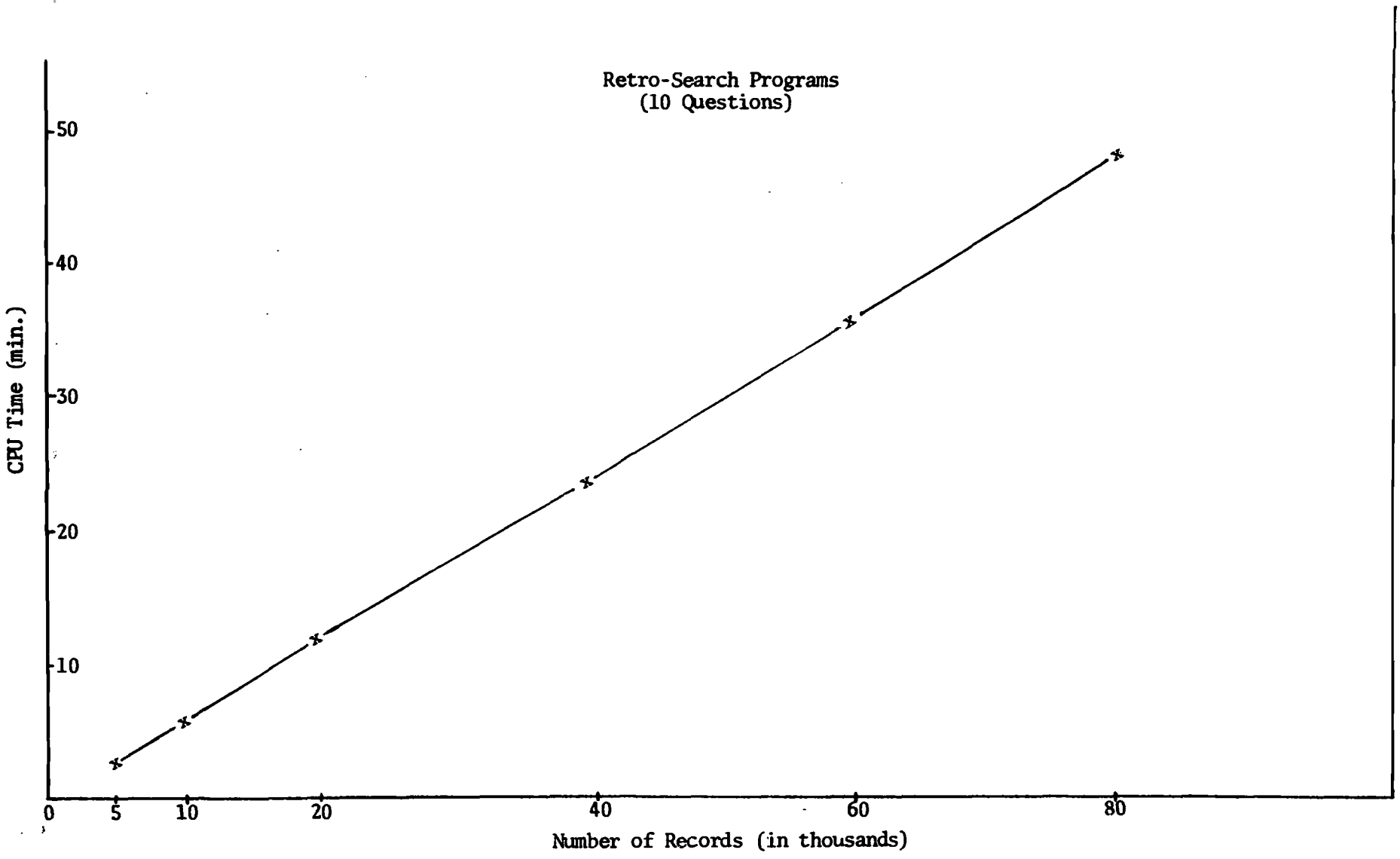


Fig. 9 Number of Records vs. CPU Time

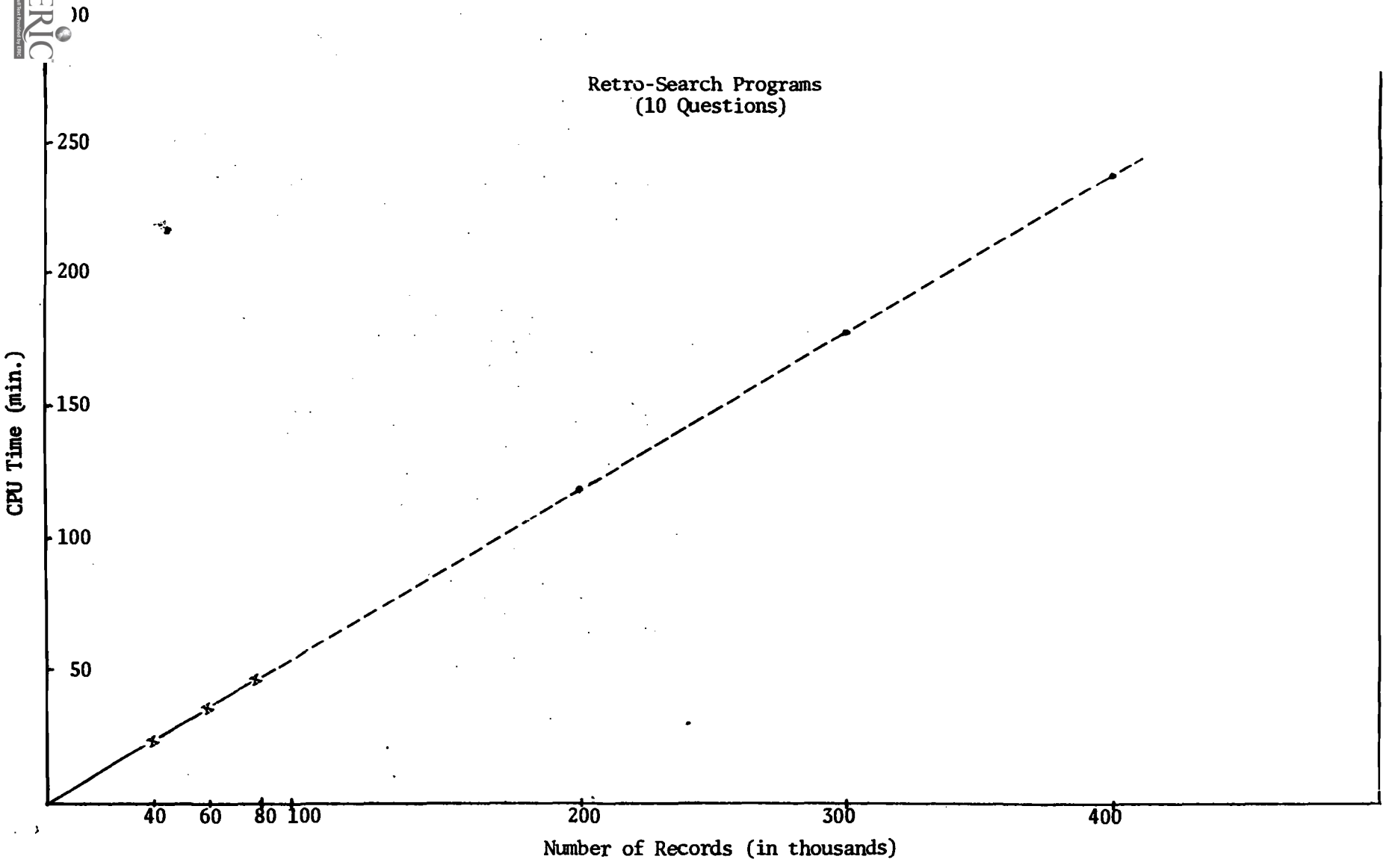


Fig. 10 Number of Records vs. CPU Time (extrapolated)

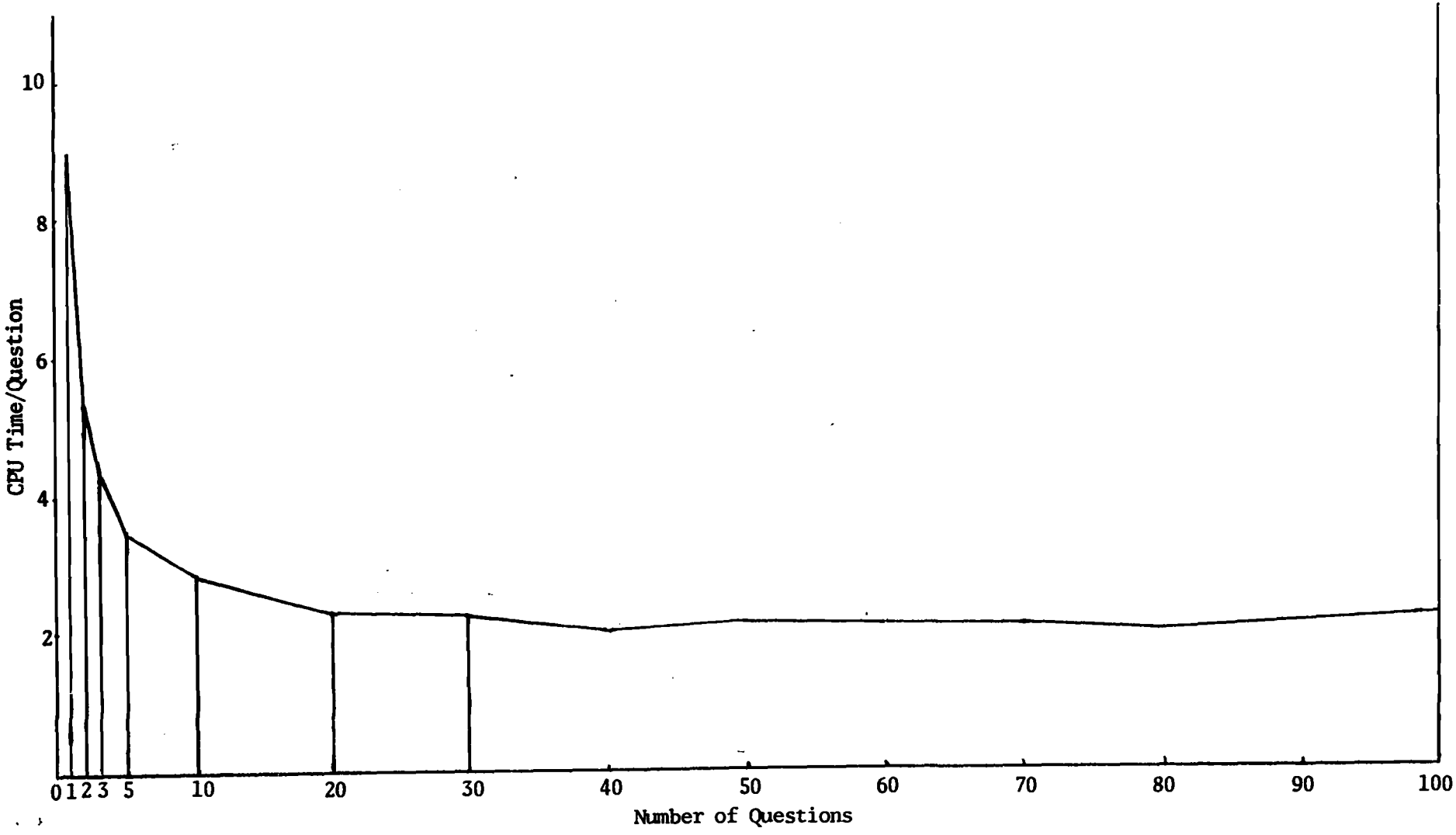


Fig. 11 CPU Time/Question (60,000 Records, 12 Hits per Question)

6.2 Cost of the Service

In this chapter we will investigate the cost of the retrospective search. Since the average increase of the COMPENDEX data base is about 60,000 records (twelve monthly tapes, each encompassing about 5,000 records) a year, we adopted this figure as the base of our calculation. The cost was computed for 5 and 50 questions representing both a small and a large batch of queries. These two calculations were done for the statistical mode also to compare it with the non-statistical.

Therefore, the following costs were assessed (Figure 12):

I	II	III	IV
Retro-Search	Retro-Search	Retro-Search	Retro-Search
Non-statistical	Statistical	Non-statistical	Statistical
5 questions	5 questions	50 questions	50 questions
60,000 records	60,000 records	60,000 records	60,000 records

Fig. 12 Cost Calculations

In calculating the Computer Job Cost, we used the following pricing structure:

1. The cost of the CPU time was calculated at $85.00 per hour.
2. Core-time cost "C" was obtained by the formula where

   $C = R \times (C_t + I_t) \times 0.20$

   R = Core requested (K)

   C_t = CPU time (hours)

   I_t = Input/Output time (hours)

3. Input/Output-time cost "I" was calculated using the formula

   $$I = \frac{(I_c \times 0.09 \text{ sec.})}{3{,}600} \times 60 = \frac{I_c \times 0.09 \text{ sec.}}{60}$$

   where I_c = Input/Output count (I/O Waits)

Total Computer Job Cost "CJC" is the sum of the component costs:

CJC = CPU + C + I

From the point of view of cost-accounting we may group the programs as follows:

1. 360 Condensed Text Edit and Edit Convert. These programs are run once for the lifetime of the data base. As an SDI service is run regularly, it is best to include this cost in the SDI cost. (We could charge it to the Retro-Search, but this would be only a guess as we do not know in advance how many questions will ever be submitted.) This way, we keep the cost of the Retro-Search lower and promote its usage, thereby enhancing the utilization of the data base. The same holds true of other costs, e.g., the data base tape cost.

2. Retro-Merge and Retro-Master Merge

   2.1 Retro-Merge provides for merging of two tapes: 360 Condensed Text and Search Text. This process is concerned with some 5,000 records each month and is performed for the retrospective module only. As the CPU time required is about 3 minutes (or $4.25), we can charge approximately $10.00 per run (in our calculations we assume one batch-run per month) as a lump sum.

   2.2 Retro-Master Merge merges old masters with the new master once a month. The old master continues to grow from month to month. We have found that the CPU time required for this program is cca 1 minute. for each 10,000 records of "New Master Totals," which is the sum of "Old Master Totals" and "Change Tape Totals." The core required is 158K. The "Input/Output Count" is roughly equal to the "New Master Totals."

3. The Question Programs

   3.1 The "Question Sort" is not performed.

3.2 The "Sorted Question Diagnostic" is considered negligible.

4. The Retro-Search Programs. They relate directly to the search respective and are included in its cost.

I. Retro-Search, Non-statistical; 5 Questions; 60,000 Records; 12 Hits per Question

A. Computer Costs

Edit Pgms:	360 Condensed Text Edit (Accounted for in the CIS service)	$000.00	$
	Edit Convert	000.00	000.00
Merge Pgms:	Retro-Merge		
	See explanation above	10.00	
	Retro-Master Merge		
	CPU time		
	60,000 records = cca 6 min.	8.50	
	I/O time	90.00	
	Core time	50.56	159.06
Question Pgms:	Retro-Question Sort not performed	000.00	
	Sorted Question Diagnostic		
	Negligible (2 sec. CPU time)	000.00	000.00
Search Pgms:	Retro-Memory Load		
	Negligible (4 sec. CPU time)	000.00	
	Retro-Search		
	CPU time = 17 min.	24.10	
	I/O time	46.50	
	Core time	10.58	
	Carried Forward	81.18	159.06

	Forwarded	$ 81.18	$159.06
Search Pgms:	Retro-Text Expansion		
	Negligible (1 sec. CPU-time)	000.00	
	Retro-Text Sort		
	Negligible (2 sec. CPU-time)	000.00	
	Retro-Print		
	CPU time = 7 sec.	0.17	
	I/O time	2.50	
	Core time	0.46	84.31
Printing:	Printing		
	5 questions with 12 hits each, makes up 60 answers. Each answer on average 23 lines equals 1,380 lines		
	$1.00 per 1,000 lines	1.38	1.38
	Total Computer Processing Costs	244.75	

B. Cost of the System (TEXT-PAC)

The system was acquired free of charge.	000.00	000.00

C. Cost of Implementation

This is not included in the cost calculation	000.00	000.00

D. Search Editing, etc.

Prompting the service, question construction, interviewing or corresponding with the user, question adjustment, coding, submitting jobs--3 hr/question		
5 x 3hr. x $5.00	75.00	75.00
		319.75

E. Keypunching-Verifying

5 questions = 6 min. = $7.00 ÷ 10	0.70	0.70
Carried Forward		$320.45

	Forwarded		$320.45
F. Material			
Data Base (tapes) } Tape Reel }	Accounted for in CIS	$000.00 000.00	
Printing Paper			
5 questions with 12 hits each = 60 answers			
3 answers cover on average 2 printed sheets			
= 40 sheets + computer data = 50 sheets of			
printing paper		0.23	
Punched Cards: 20 lines x 5 questions = 100 cards		0.11	0.34
G. Handling, Mailing, etc.			
2% of the D. costs		1.50	1.50
H. Other Overhead Cost			
This is included in A.			000.00
Total Cost per 5 questions			322.29

1 Question = $322.29 ÷ 5 = $64.46

II. Retro-Search, Statistical; 5 Questions; 60,000 Records; 12 Hits/Question

There are two additional programs as compared with the non-statistical run.

1. Retro Answer Sort			
CPU time		0.10	
I/O time		0.29	
Core time		0.09	
2. Retro-Statistical			
CPU time		0.07	
I/O time		0.36	
Core time		0.06	
Total in addition to "non-statistical"		0.97	0.97
Retro-Search non-statistical			322.29

Total Statistical (5 questions; 60,000 records) $323.26

1 Question $64.65

III. Retro-Search, Non-statistical; 50 Questions; 60,000 Records; 12 Hits/Question

A. Computer Cost

Edit Pgms:	360 Condensed Text Edit	$000.00	
		Accounted for in the CIS service	
	Edit Convert	000.00	
Merge Pgms:	Retro-Merge		
	See explanation above	10.00	
	Retro-Master Merge		
	CPU time		
	60,000 records = cca 6 min.	8.50	
	I/O time	90.00	
	Core time	50.56	159.06
Question Pgms:	Retro-Question Sort not performed	000.00	
	Sorted Question Diagnostic		
	Negligible	000.00	000.00
Search Pgms:	Retro-Memory Load		
	CPU time (13 sec.)	0.31	
	I/O time	1.33	
	Core time	0.33	
	Retro-Search		
	CPU time = 108 min.	153.00	
	I/O time	47.25	
	Core time	41.40	
	Retro-Text Expansion		
	CPU time	0.07	
	Carried Forward	243.69	159.06

			Forwarded $243.69	$159.06
		I/O time	0.23	
		Core time	0.05	
		Retro-Text Sort		
		CPU time	0.68	
		I/O time	4.20	
		Core time	1.12	
		Retro-Print		
		CPU time	0.09	
		I/O time	24.00	
		Core time	4.17	278.23
	Printing:	Printing		
		50 Questions with 12 hits each, equals 600 answers. Each answer contains an average of 23 lines = 13,800 lines $1.00 per 1,000 lines	13.80	13.80
		Total Computer Processing Cost	451.09	
B.	Cost of the System (TEXT-PAC)			
	The system was acquired free of charge		000.00	000.00
C.	Cost of Implementation			
	This is not included in the cost calculation		000.00	000.00
D.	Search Editing, etc.			
	Promoting the service, question construction, interviewing or corresponding with users, question adjustment, coding, submitting jobs 3 hr./question			
		50q x 3 hr. x $5.00	750.00	750.00
E.	Keypunching-Verifying			
	1 hour (on average)		7.00	7.00
		Carried Forward		1,208.09

	Forwarded		$1,208.09
F.	Material		
	Data Base (tapes) } Accounted for		
	Tape Reels } in CIS		
	Printing Paper		
	50 questions with 12 hits each = 600 answers		
	3 answers per 2 sheets = 400 sheets		
	(15" x 8.5") $4.50/1,000 400 = $1.80	1.80	
	Punched Cards: 20 lines x 50 questions = 1,000	1.10	2.90
G.	Handling, Mailing, etc.		
	2% of the D. costs	15.00	15.00
H.	Other Overhead		
	This is included in A.	000.00	000.00
	Total Cost per 50 questions		1,225.99
	1 Question = $1,225.99 ÷ 50 = $24.52		

IV. Retro-Search, Statistical; 50 Questions; 60,000 Records; 12 Hits/Question

There are two additional programs as compared with the non-statistical run.

1.	Retro-Answer Sort		
	CPU time	0.16	
	I/O time	0.86	
	Core time	0.25	
2.	Retro-Statistical		
	CPU time	0.30	
	I/O time	4.51	
	Core time	0.72	
	Total in addition to "non-statistical"	6.80	6.80
	Retro-Search non-statistical		1,225.99
	Total Statistical (50 questions; 60,000 records)		1,232.79
	1 Question $24.66		

From these cost calculations several conclusions may be drawn. First of all, we can infer that the statistical option should be used wherever needed because of its merits and low additional cost (Figure 13):

1 Question

Out of Five		Out of Fifty	
Non-statistical	Statistical	Non-statistical	Statistical
$64.46	$64.65	$24.52	$24.66

Fig. 13 Statistical/Non-statistical

Secondly, questions should be run in optimum batches. Whereas the size of a batch cannot influence the question-dependant costs under D (Search Editing, etc.), E (Keypunching), G (Handling, Mailing, etc.), and partly F (Material), it will have a marked effect on the total and computer costs as may be seen from the tables above. We have already stated that in our example the CPU time required to run 1 question is 8.5 minutes as compared with 2 minutes per question when processing a 40-question batch. The optimum search time sets in at 20 questions and extends up to the other limiting factor which is the capability to process one "memory load" of questions at one time: one memory load is approximately 100 questions (or slightly above, depending on the size of questions). If more than one memory load of questions are to be processed, two or more runs will be necessary.

Yet this optimum range of questions to be processed at one time (20 through 100) has another restrictive condition, namely the number of hits. The maximum number of hits which can be handled by the "Retrospective Text Sort" program is 6,000. A larger number of hits can be accommodated by using the IBM 360/OS Sort Program. An excessive amount of hits, however, prevents other users from running their jobs for hours. It seems, therefore, reasonable to recommend, at least on our configuration, to set the limit of 6,000 hits and run 20 questions with an average of 300 hits, or 30 questions with 200 hits each, and so forth.

Also a batch of questions with both high and lower requested number of hits will certainly occur. This way, other users will be able to use the core, disks or tape for their jobs on the system.

There are four ways to keep the number of hits in reasonable limits: (1) to reduce the number of questions in the batch; (2) to split the data base into subsets; (3) to specify in the Header card a lower number of answers required; (4) to use search logic to obtain the desired effect. Approach (1) will necessitate more runs with higher costs per question. The same applies to solution (2)if we split the data base into one-year data bases. If we specify a lower number of "wanted hits" (3), then "the wanted number" might be in some cases filled with the oldest information from the beginning of the data base and the user would miss the most desired recent information.

For this reason we recommend approach (4) using the search logic to achieve the desired effect: to get the number of hits we want as a relevance/recall trade-off.

After a couple of years the size of the data base would make the search too lengthy and costly. As already mentioned the expected yearly growth is 60,000-70,000 records. After five years the data base would represent 300,000-350,000 records on 30-35 tapes. As our graph (Figure 10) indicates this would require 180-210 minutes of search time for 10 questions with a small number of hits. The most appropriate solution to this problem seems to be to subdivide the data base into a series of subject areas. This would enable us to confine the search to a data-base of a limited size and obviate searching in its irrelevant regions. There is a catch in it, too, since we cannot conduct the search for a question in tape A, and for another question in the tape B, in the same batch of questions. We would have to run a batch of questions in related areas each time. However, with a vast amount of records the advantage of processing a small data base would make up for the necessity to run small batches of questions.

The Card-Alert Codes of COMPENDEX would help in creating subsets.

For example, after three years of operation, we would have some 180,000 records. At this time it would be practical to subdivide it into:

1. Civil--Environmental--Geological--Bioengineering
2. Mining--Metals--Petroleum--Fuel Engineering
3. Mechanical--Automotive--Nuclear--Aerospace Engineering
4. Electrical--Electronics--Control Engineering
5. Chemical--Agricultural--Food Engineering
6. Industrial Engineering--Management--Mathematics--Physics--Instruments

Instead of handling 18 tapes in a search, one would have to process approximately 3 of them, or 6 if the question would be expected to get response in two of the subsets specified above. After, say, two more years further splitting would take place separating e.g., aerospace engineering in a self-contained subject-field subset, and so on.

6.3 Cost/Benefit

The question, which is always asked, is whether the cost of a service is justified by the benefits from the service.

Assume we have processed a question along with others in a batch of 50 against one year's data base of 60,000 records. The cost of this search has been $24.66 (or $64.46 in a 5-question batch) with the statistical option. Most of the information services are subsidized in some way or other, so the actual price to the user would be lower.

If our user has to cope with his information problem using hard copies of an abstract journal, he obviously does not have to scan all of the 60,000 abstracts, but rather approximately 1/10 of the abstracts, in some cases more, in others less. If he goes through 1,000 abstracts he probably would scan six of them in one minute. Getting through 6,000 abstracts would reduce the efficiency of scanning to four per minute. This literature search would take 25 hours and cost $250, if we charge only the research worker's salary and disregard the value he could generate if he were freed for his special work. This would represent a multiple of this amount. If he subscribes to some file card information service, his recall will be lower than in full text searching and the price is to be added to the cost of personal searching.

Frequently, however, a literature search is not done and this

does not mean that the amount of $250 is saved. Rather, some work already done elsewhere is duplicated, other people's patent rights are infringed and the work itself is not done at the level it might have been had the literature been searched.

This once again substantiates the fact that experimenting in the literature is cheaper than experimenting in a laboratory. It also proves that some organizations could increase their capacity by as much as one third by using professional information services.

6.4 Principles of Pricing

The cost per question is increasing with the number of records, decreases with the increasing number of questions, and increases with the number of hits. Logically, the user should be paying more for searching a larger data base. This could be achieved by performing a search in, say, the last 24 months, at a standard price and, on demand, by conducting a search in the "historical" tapes at an additional price proportionate to the size of data base. This historical data base could be, as outlined above, split into subject areas, and this would mean decreased costs.

On the other hand, the user should not be billed more because his question was processed in a small batch, unless he insisted on a prompt search.

As the number of hits affects the cost of the searching [the difference between a low number of hits (12) and a high number of hits (average of 1,400) was nearly 100 per cent more search time, for the same number of questions (20) and the same data base (115,000 records)], users should pay some additional fee for more hits. Indirectly, wanting many hits in any question will require running a smaller batch and cause higher costs.

Of course, the price should also reflect the size of the question, either in words or in search expressions.

In practice, users should be told of an average cost calculated above under the given conditions. They should agree that the actual cost will be computed after it has been processed, as outlined above.

We might also want to prepare a rough estimate of the cost generally to provide ourselves with some pricing mechanism. Our estimate is based

on the assumption that the average number of questions processed in the monthly batch will be 25 (see Figure 14). We have calculated the values M and N. An analysis of these values indicates that the cost of the retrospective searches C has two major components: editing E and processing P:

$$C = E + P \quad \text{or}$$

$$Cp = (Np \times Hp \times Wh + R + S) \div Np$$

where Cp = Cost per Profile
Np = Number of Profiles
Hp = Hours per Profile (Editing)
Wh = Wage per Hour
R = Retro-Master Merge
S = Searching Programs

Other component items play a minor role. The most significant of the Search programs is the Retro-Search (which can be used for a rough estimate).

As the Computer Processing Costs pare directly proportional to the number of records and the Search Editing Costs are directly proportional to the number of questions, we can estimate the cost per question by approximation as shown in the following table (Figure 15).

If we draw the lines between the calculated values M and N, and the estimated values R and S (see Figure 14 and Figure 15), we can find for 25 questions the prints A, B, and C giving the rough costs for an accepted average number of questions:

A ≈ $ 45.00 (1 year data base)
B ≈ 77.50 (2 years data base)
C ≈ 110.00 (3 years data base)

These estimated costs apply up to an average number of question (profile) words, i.e., 40. Each word above this limit should be charged an additional $1.00.

As these costs represent a low number of hits (and the hits affect the search time), there should be an additional charge for an excessive number of hits:

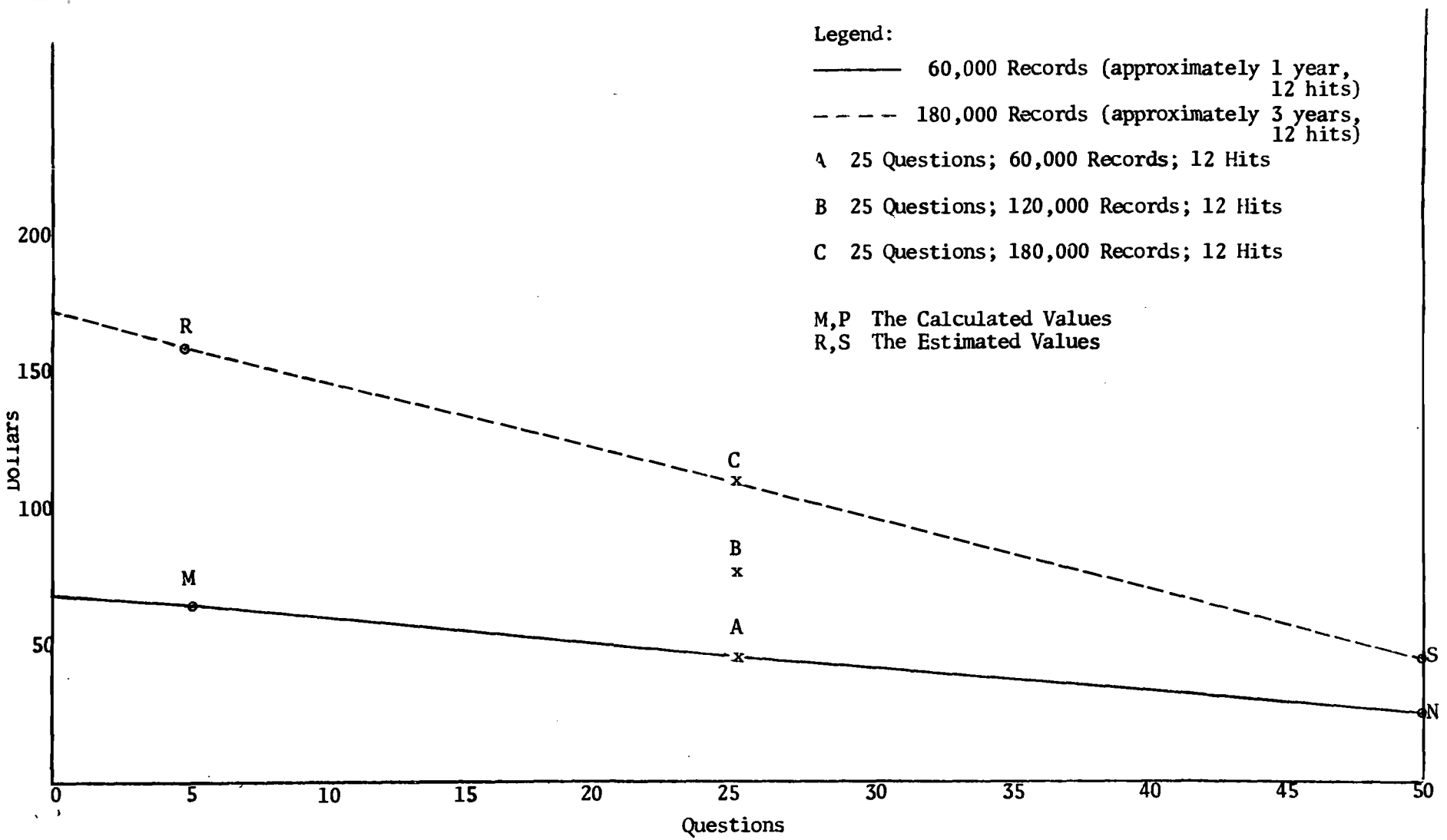


Fig. 14 Rough Estimate of Cost Per Question

Conditions / Costs	5 Questions, 12 Hits per Question	
	60,000 Records (calculated)	180,000 Records (estimated)
	$	$
Computer Cost	245	735*
Search Editing Cost	75	75**
Diverse	3	
Total Cost	323	810
Cost per Profile	65 (Value M)	162 (Values R)

*Three times 245
**Remains unchanged

Conditions / Costs	50 Questions, 12 Hits per Question	
	60,000 Records (calculated)	180,000 Records (estimated)
	$	$
Computer Cost	458	1,374*
Search Editing Cost	750	750**
Diverse	25	
Total Cost	1,233	2,224
Cost per Profile	25 (Value P)	44 (Value S)

*Three times 458
**Remains unchanged

Fig. 15 Calculated and Estimated Values for 1, 2, and 3 Years' Data Base Searching

If there are >	50 hits	each hit 1¢
>	100 hits	2¢
>	200 hits	3¢
>	300 hits	4¢

7. CIS IN RETRO-SEARCH MODULE

As the Retrospective-Search module has the "statistical option" indicating the matched words by a particular document, and the CIS module does not, we had to solve the following dilemma: either (1) to "transplant" this option to the CIS section, or (2) to use the Retrospective-Search section to process the CIS profiles. The first alternative would entail study and reprogramming, but would not necessitate a change in the Header cards and would leave the output (the double cards) unchanged. The second alternative was chosen because the profiles can be run after minor formal changes (see the CIS profile form and Retro-Search question form for more details) with the limitations as they were outlined for the Retro-Search: only one memory load of profiles can be run at one time; output is on stack printing paper; maximum 6,000 hits are recommended.

The costs of running 100 questions (in CIS called profiles) against 5,000 documents, with the statistical option, producing 5 hits per question, are analyzed below. The times for the 360 Condensed Text Edit and Edit Convert were taken from our COMPENDEX/TEXT-PAC/CIS Report where the database examined contained 4,848 records.

Retro-Search, Statistical; 100 Questions;
5,000 Records; 5 Hits/Question

A. Computer Cost

Edit Pgms: 360 Condensed Text Edit
Data taken from the "COMPENDEX/TEXT-PAC/CIS" Report
CPU time 41.23 min.
Core required 100K
I/O counts 42

	CPU time cost	$ 58.41	$
	I/O time cost	0.06	
	Core time cost	13.76	72.23
	Edit Convert		
	Data taken from the "COMPENDEX/		
	TEXT-PAC/CIS" Report		
	CPU time 25.33 min.		
	Core required 128K		
	I/O count 55		
	CPU time cost	36.17	
	I/O time cost	0.08	
	Core time cost	10.91	47.16
Merge Pgms:	Retro-Merge		
	See explanation above	10.00	10.00
	Retro-Master Merge		
	5,000 records = 0.5 min. CPU time		
	CPU time cost	0.71	
	I/O time cost	7.50	
	Core time cost	4.21	12.42
Question Pgms:	Retro-Question Sort not performed	000.00	
	Retro-Question Diagnostic		
	Negligible (9 sec. CPU)	000.00	000.00
Search Pgms:	Retro-Memory Load		
	CPU time 43 sec.		
	Core required 106K		
	I/O count 2,182		
	CPU time cost	1.02	
	I/O time cost	3.27	
	Core time cost	1.41	5.70
	Retro-Search		
	CPU time 21 min. 13 sec.		
	Core required 132K		
	I/O count 3,120		
	CPU time cost	30.06	
	I/O time cost	4.68	
	Core time cost	11.40	46.14
	Carried Forward		$193.65

Forwarded		$193.65
Retro-Answer Sort		
CPU time 7 sec.		
Core required 76K		
I/O count 491		
CPU time cost	0.17	
I/O time cost	0.74	
Core time cost	0.24	1.15
Retro-Statistical		
CPU time 17 sec.		
Core required 86K		
I/O count 3,984		
CPU time cost	0.40	
I/O time cost	5.98	
Core time cost	1.80	8.18
Retro-Text Expansion		
CPU time 2 sec.		
Core required 50K		
I/O counts 96		
Negligible	000.00	000.00
Retro-Text Sort		
CPU time 18 sec.		
Core required 72K		
I/O count 2,194		
CPU time cost	0.43	
I/O time cost	3.29	
Core time cost	0.86	4.58
Retro-Print		
CPU time 2 sec.		
Core required 52K		
I/O count 275		
CPU time cost	0.05	
I/O time cost	0.38	
Core time cost	0.07	0.50
Carried Forward		$208.06

	Forwarded	$	$208.06
Printing:	Printing 100 profiles with 5 hits each equals 500 answers. Each answer consists of an average of 23 lines = 12,500 lines $1.00 per 1,000 lines	12.00	12.00
	Total Computer Processing Costs = $220.06		
B.	Cost of the System (TEXT-PAC)		
	The system was acquired free of charge.	000.00	000.00
C.	Cost of Implementation		
	This is not included in the cost calculation	000.00	000.00
D.	Search Editing, etc.		
	Promoting the service, profile construction, interviewing or corresponding with users, profile adjustment coding, submitting jobs	400.00	400.00
E.	Keypunching-Verifying		
	1 hour (an average)	7.00	7.00
F.	Material		
	Data Base Tapes (one monthly tape @ $500.00)	500.00	
	One reel @ $25.00	25.00	
	Printing Paper		
	100 profiles = 500 answers		
	3 answers per 2 printing sheets		
	Answers + statistical data + other data =		
	500 sheets (15" x 8.5") 1,000 sheets @ $4.50	2.25	
	Punch Cards, cca 20 lines per profile 2,000 lines = 2,000 cards	2.20	529.45
	Carried Forward		1,156.51

	Forwarded	$	$1,156.51
G. Handling, Mailing			
2% of the D. cost		30.00	30.00
H. Other Overhead			
This is included in A.		000.00	000.00
Total Cost per 100 profiles/month			$1,186.51

1 profile = $11.87/month

In this total cost of the monthly processing of 100 profiles ($1,186.51), the proportion of the individual most significant cost items may be singled out as illustrated (Figure 16).

As may be seen from the diagram, the most significant cost item is represented by the data-base tapes with reels which amount as high as 44.3 per cent of the total. This illustrates also the way to go if we plan to enhance the economy of the service: to process as many profiles as possible (with physical limitations in view) to keep the proportion of this cost per profile low. Further, the economy of the CIS service can be improved by retrospective searches which should be given wide publicity. Only the multiple use of this data base can make it economically viable. As it is a fixed cost, its proportion per profile is decreasing with the rising number of profiles.

Search Editing represents a proportional cost which increases directly with the number of profiles. Seemingly, we can get more out of a monthly salary if we divide it by a higher number of profiles. This is a wrong approach, though, as it affects the quality. There is a certain limit imposed on the capacity of a search editor and after that we need additional search editors which, in turn, increases the costs. An ideal solution seems to be processing up to 100 profiles in the CIS, each of them with a life-span of at least 5-10 months. The rest of the search editor's capacity ought to be directed to the retrospective searching (at least 20 searches per monthly run).

The computer processing is a rather surprisingly low percentage of the total cost. Some of its components are proportional cost (e.g., editing cost rising with the data-base and profiling cost rising with

the number of profiles), search time represents a cost proportional to the size of data-base (Figure 10), and to the number of questions (see Figure 8).

The cost calculated in this CIS run is considerably less than in (2) of the previous report. The computer cost is lower mainly because of the lower CPU rate; also the search time is less (21 minutes for 100 profiles in the retrospective module, related to 28 in the CIS module.) According to the graph in Figure 36 of the report (2) the search time for 100 profiles would be 40 minutes. Also, no reserve is taken for the dictionary (profiles will be improved by means of the statistical printout), no consulting is included, salaries are lower in the production runs compared to the developmental stage. The output is also cheaper on the paper as compared with the double cards.

The users of COMPENDEX-CIS(SDI) service would receive printed sheets instead of cards. They would have the choice: (1) to receive the statistical data regarding hits and adjust the profiles themselves (or give suggestions as to changes), (2) leave the adjusting of profiles to search editors who would keep the statistical printout for this purpose; in this case the user would send all completely irrelevant abstracts back to the search editor.

Modifying the print program to print the answers on the double cards would be relatively easy, should the users prefer it.

As far as feedback is concerned, we suggest that the users be asked only to send back the completely irrelevant abstracts.

8. CONCLUSIONS

Retrospective Searching in the TEXT-PAC System can be defined as computer matching of a machine-readable data base prepared as a result of human abstracting and indexing, against one or more questions intellectually prepared and translated into the system language. The entire record is scanned for occurrence of the question words and logic. The "hits" are obtained in the form of a computer printout. The "statistical" option may be required which indicates the words and logic responsible for matches. The mode of computer processing is local batch.

the number of profiles), search time represents a cost proportional to the size of data-base (Figure 10), and to the number of questions (see Figure 8).

The cost calculated in this CIS run is considerably less than in (2) of the previous report. The computer cost is lower mainly because of the lower CPU rate; also the search time is less (21 minutes for 100 profiles in the retrospective module, related to 28 in the CIS module.) According to the graph in Figure 36 of the report (2) the search time for 100 profiles would be 40 minutes. Also, no reserve is taken for the dictionary (profiles will be improved by means of the statistical printout), no consulting is included, salaries are lower in the production runs compared to the developmental stage. The output is also cheaper on the paper as compared with the double cards.

The users of COMPENDEX-CIS(SDI) service would receive printed sheets instead of cards. They would have the choice: (1) to receive the statistical data regarding hits and adjust the profiles themselves (or give suggestions as to changes), (2) leave the adjusting of profiles to search editors who would keep the statistical printout for this purpose; in this case the user would send all completely irrelevant abstracts back to the search editor.

Modifying the print program to print the answers on the double cards would be relatively easy, should the users prefer it.

As far as feedback is concerned, we suggest that the users be asked only to send back the completely irrelevant abstracts.

8. CONCLUSIONS

Retrospective Searching in the TEXT-PAC System can be defined as computer matching of a machine-readable data base prepared as a result of human abstracting and indexing, against one or more questions intellectually prepared and translated into the system language. The entire record is scanned for occurrence of the question words and logic. The "hits" are obtained in the form of a computer printout. The "statistical" option may be required which indicates the words and logic responsible for matches. The mode of computer processing is local batch.

The COMPENDEX data base is available commencing January, 1969 and the yearly growth is expected to be 60,000-70,000 records, or six to seven tapes. The data base has proven to have a good mega-relevance to all of the areas of engineering. The system can operate over a wide range of relevance and recall values.

It has been shown that the CPU-time of the search programs is influenced by the number of questions, by the number of data-base records and hits. We have found that one-question run requires as much as 8.5 minutes of the CPU time, whereas with a 40-question batch only two minutes per question are needed. The optimum search time sets in at 20 questions and extends up to the "memory load" or approximately 100 questions which can be processed in one run. The maximum number of matches processed in one run should be about 6,000, otherwise the standard utility sort program has to be used. An excessive amount of hits may inconvenience other users of the computer system by occupying the auxiliary storage devices, so 6,000 hits is a practical upper limit.

The statistical option should be used because of its merits and low additional cost. The cost of one question in a five-question batch is \$64.46 (statistical \$64.65), and it drops to \$24.52 (statistical \$24.66) for one question out of fifty; this applies to searching 60,000 records and 12 hits per question. These figures illustrate the effect of running the optimum size batches (20-100 questions).

It is suggested that the CIS service or SDI (Selective Dissemination of Information) be also run in the Retrospective Search module. This would enable us, with the statistical printout at hand, to adjust the profiles accordingly. We regard the statistical option as even more significant in the SDI service in view of the dynamic character of profiles. The costs of searching are reasonable. (One profile out of one hundred costs \$11.87 per month, with five received answers.) Since the cost of the data base is the most expense, a better economy can be achieved by greater use of it.

The SDI feedback procedure could be further simplified; the users would be expected to send back only the completely irrelevant abstracts. The profiles could be corrected by means of the statistical printout and

the irrelevant abstracts.

In view of the substantial yearly data base increase it is suggested that the last one or two years' data base be searched after simple merging, but the "historical" data base should be presorted to make up subject-area tapes. The Card-Alert Codes of Engineering Index would serve this purpose. Through this subsetting, the data base searched could be maintained at a reasonable size.

Users should be charged depending on the size of their question (number of words or search expressions), the size of the data base they specify in the Header card, and the number of hits they receive. They should be advised of the costs in the above examples and they should agree to pay actual costs computed after the run.

The submission form has been prepared as well as the user for retrospective searches and they are available on request.

9. REFERENCES

(1) Kanfman, S. TEXT-PAC, S1360 Normal Text Information Processing, Retrieval and Current Information System (360.D.06.7.020). Armonk, N. Y., 10504: ITIRC, IBM Corporate Headquarters, 1968.

(2) Standera, O. R. COMPENDEX/TEXT-PAC(CIS) Project Report. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(3) _____. COMPENDEX Profiling Guide. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(4) _____. COMPENDEX Retrospective Search Instructions. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(5) _____. Profile Adjustment Manual. Information Systems and Services Division. Calgary: The University of Calgary, 1970.

(6) _____. TEXT-PAC Input, Bulletin and Indexes. Information Systems and Services Division. Calgary: The University of Calgary, 1971.

(7) Carroll, K. D. Survey of Scientific-Technical Tape Services. Information Division. New York: American Institute of Physics, 1970.

(8) Cuadra, C. A. Annual Review of Information Science and Technology. Vol. 5. American Society for Information Science. Chicago: Encyclopaedia Britannica Inc., 1970.

(9) Cohan, L. (ed.). Directory of Computerized Information in Science and Technology. New York: Science Associates/International, Inc., 1968.

(10) PIE-Publications Indexed for Engineering. New York: Engineering Index, Inc., 1970.

(11) SHE-Subject Headings for Engineering. New York: Engineering Index, Inc., 1970.

(12) Card-A-Lert-Selective Information Service. New York: Engineering Index, Inc., 1970.

(13) COMPENDEX-Computerized Engineering Index. New York: Engineering Index, Inc., 1970.